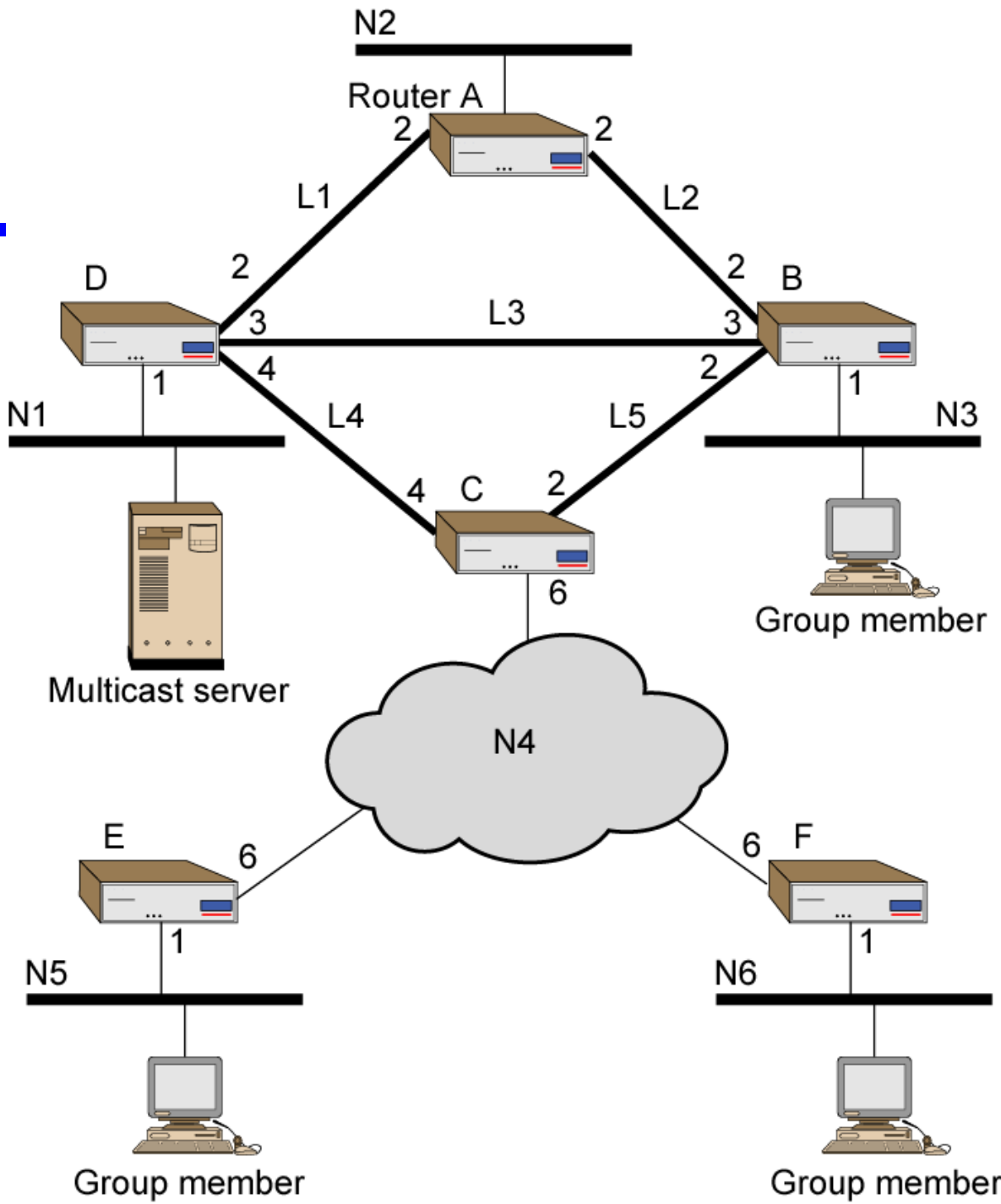
William Stallings
Data and Computer Communications
7th Edition

Chapter 19
Internetwork Protocols

Multicasting

- Addresses that refer to group of hosts on one or more networks
- Uses
  - Multimedia “broadcast”
  - Teleconferencing
  - Database
  - Distributed computing
  - Real time workgroups

Example Config



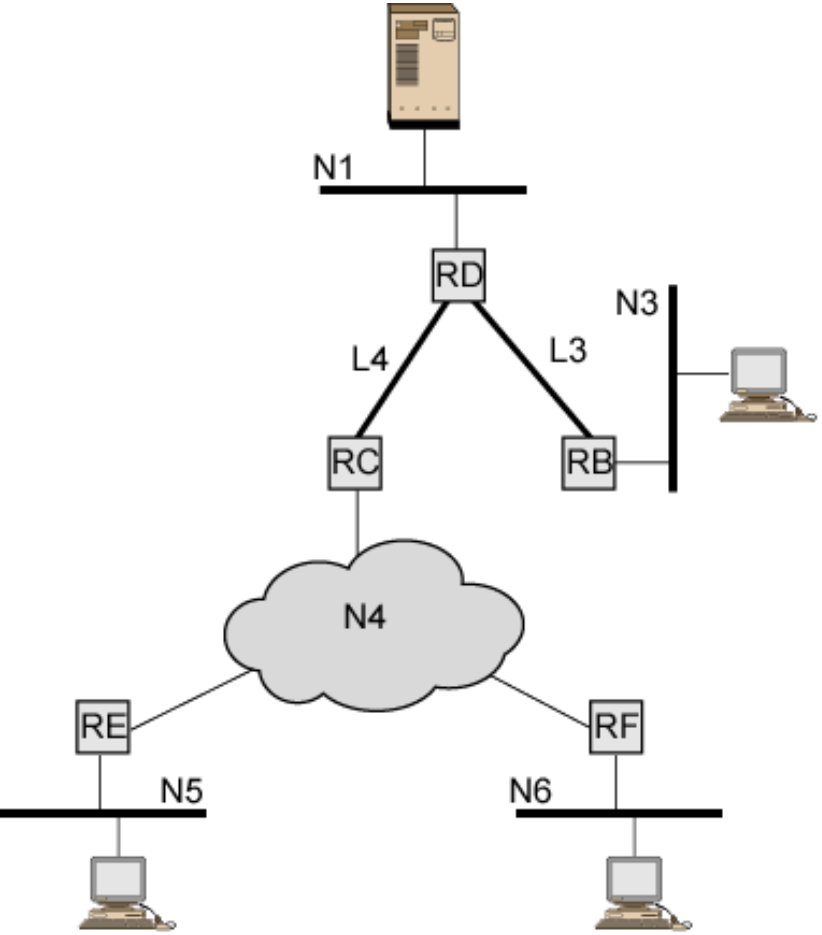
Broadcast and Multiple Unicast

- Broadcast a copy of packet to each network
  - —Requires 13 copies of packet
- Multiple Unicast
  - —Send packet only to networks that have hosts in group
  - —11 packets

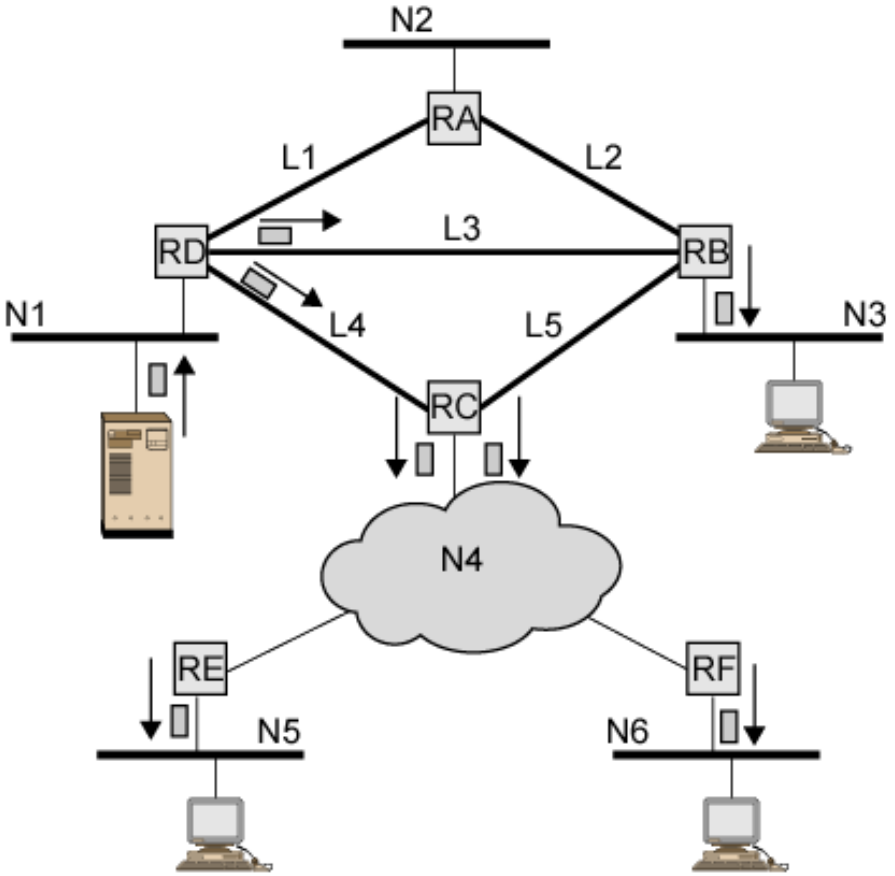
True Multicast

- Determine least cost path to each network that has host in group
  - Gives spanning tree configuration containing networks with group members
- Transmit single packet along spanning tree
- Routers replicate packets at branch points of spanning tree
- 8 packets required

Multicast Example



(a) Spanning tree from source to multicast group



(b) Packets generated for multicast transmission

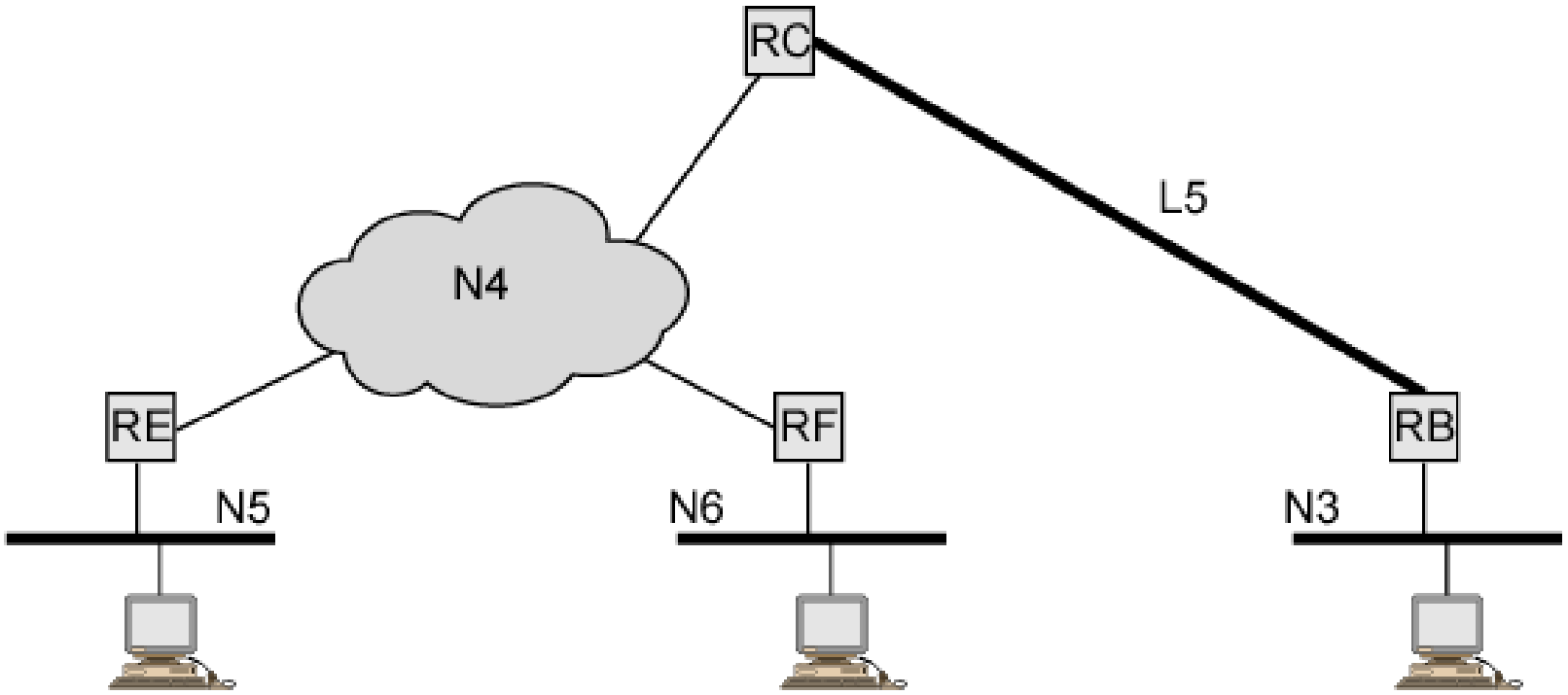
Requirements for Multicasting (1)

- Router may have to forward more than one copy of packet
- Convention needed to identify multicast addresses
  - IPv4 - Class D - start 1110
  - IPv6 - 8 bit prefix, all 1, 4 bit flags field, 4 bit scope field, 112 bit group identifier
- Nodes must translate between IP multicast addresses and list of networks containing group members
- Router must translate between IP multicast address and network multicast address

Requirements for Multicasting (2)

- Mechanism required for hosts to join and leave multicast group
- Routers must exchange info
  - —Which networks include members of given group
  - —Sufficient info to work out shortest path to each network
  - —Routing algorithm to work out shortest path
  - —Routers must determine routing paths based on source and destination addresses

Spanning Tree from Router C to Multicast Group



Internet Group Management Protocol (IGMP)

- RFC 3376
- Host and router exchange of multicast group info
- Use broadcast LAN to transfer info among multiple hosts and routers

Principle Operations

- Hosts send messages to routers to subscribe to and unsubscribe from multicast group
  - —Group defined by multicast address
- Routers check which multicast groups of interest to which hosts
- IGMP currently version 3
- IGMPv1
  - —Hosts could join group
  - —Routers used timer to unsubscribe members

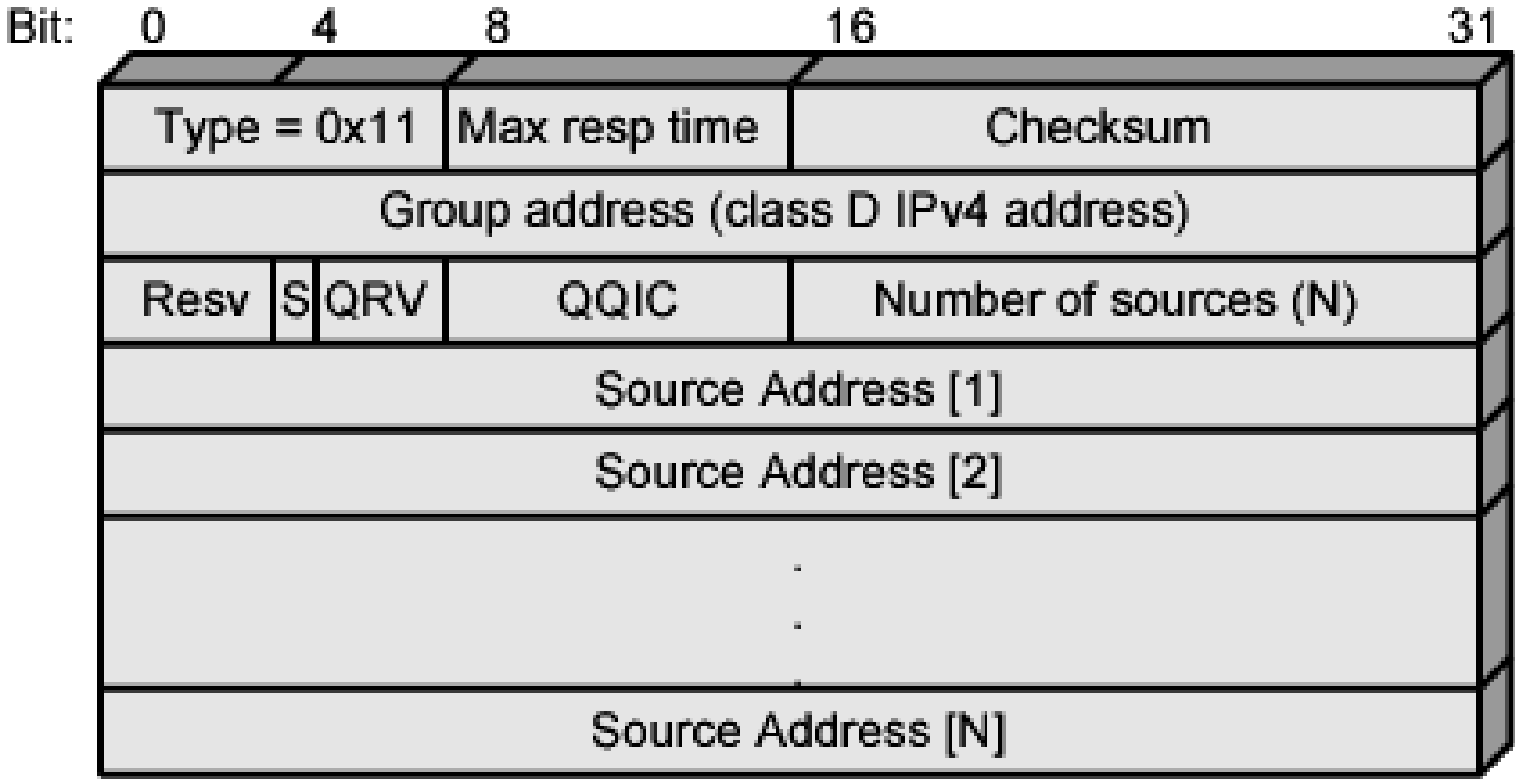
Operation of IGMPv1 & v2

- Receivers have to subscribe to groups
- Sources do not have to subscribe to groups
- Any host can send traffic to any multicast group
- Problems:
  - Spamming of multicast groups
  - Even if application level filters drop unwanted packets, they consume valuable resources
  - Establishment of distribution trees is problematic
  - Location of sources is not known
  - Finding globally unique multicast addresses difficult

IGMP v3

- Allows hosts to specify list from which they want to receive traffic
  - —Traffic from other hosts blocked at routers
- Allows hosts to block packets from sources that send unwanted traffic

IGMP Message Formats
Membership Query



(a) Membership query message

Membership Query

- Sent by multicast router
- General query
  - —Which groups have members on attached network
- Group-specific query
  - —Does group have members on an attached network
- Group-and-source specific query
  - —Do attached device want packets sent to specified multicast address
  - —From any of specified list of sources

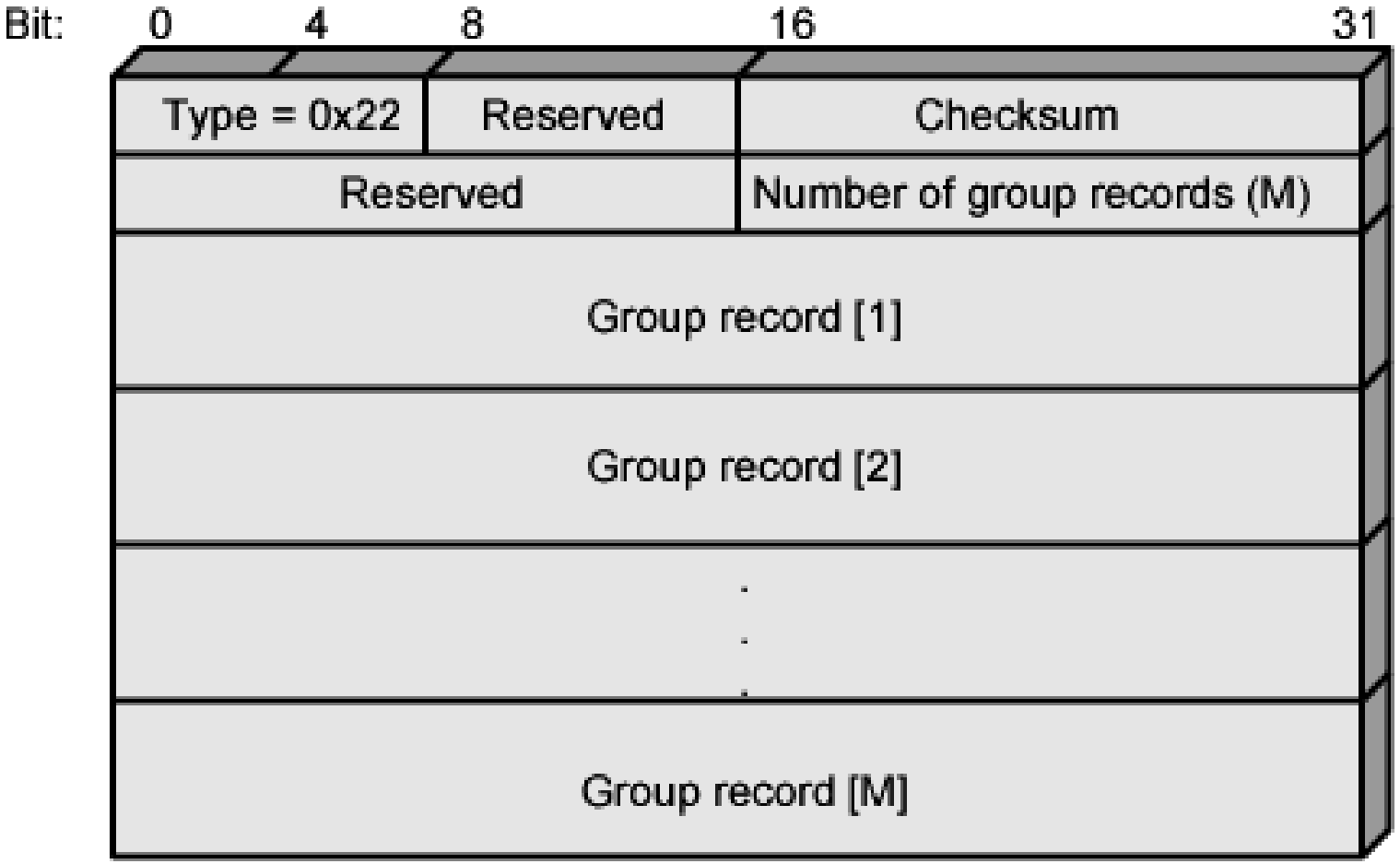
Membership Query Fields (1)

- Type
- Max Response Time
  - Max time before sending report in units of 1/10 second
- Checksum
  - Same algorithm as IPv4
- Group Address
  - Zero for general query message
  - Multicast group address for group-specific or group-and-source
- S Flag
  - 1 indicates that receiving routers should suppress normal timer updates done on hearing query

Membership Query Fields (2)

- QRV (querier's robustness variable)
  - RV value used by sender of query
  - Routers adopt value from most recently received query
  - Unless RV was zero, when default or statically configured value used
  - RV dictates number of retransmissions to assure report not missed
- QQIC (querier's querier interval code)
  - QI value used by querier
  - Timer for sending multiple queries
  - Routers not current querier adopt most recently received QI
  - Unless QI was zero, when default QI value used
- Number of Sources
- Source addresses
  - One 32 bit unicast address for each source

IGMP Message Formats
Membership Report

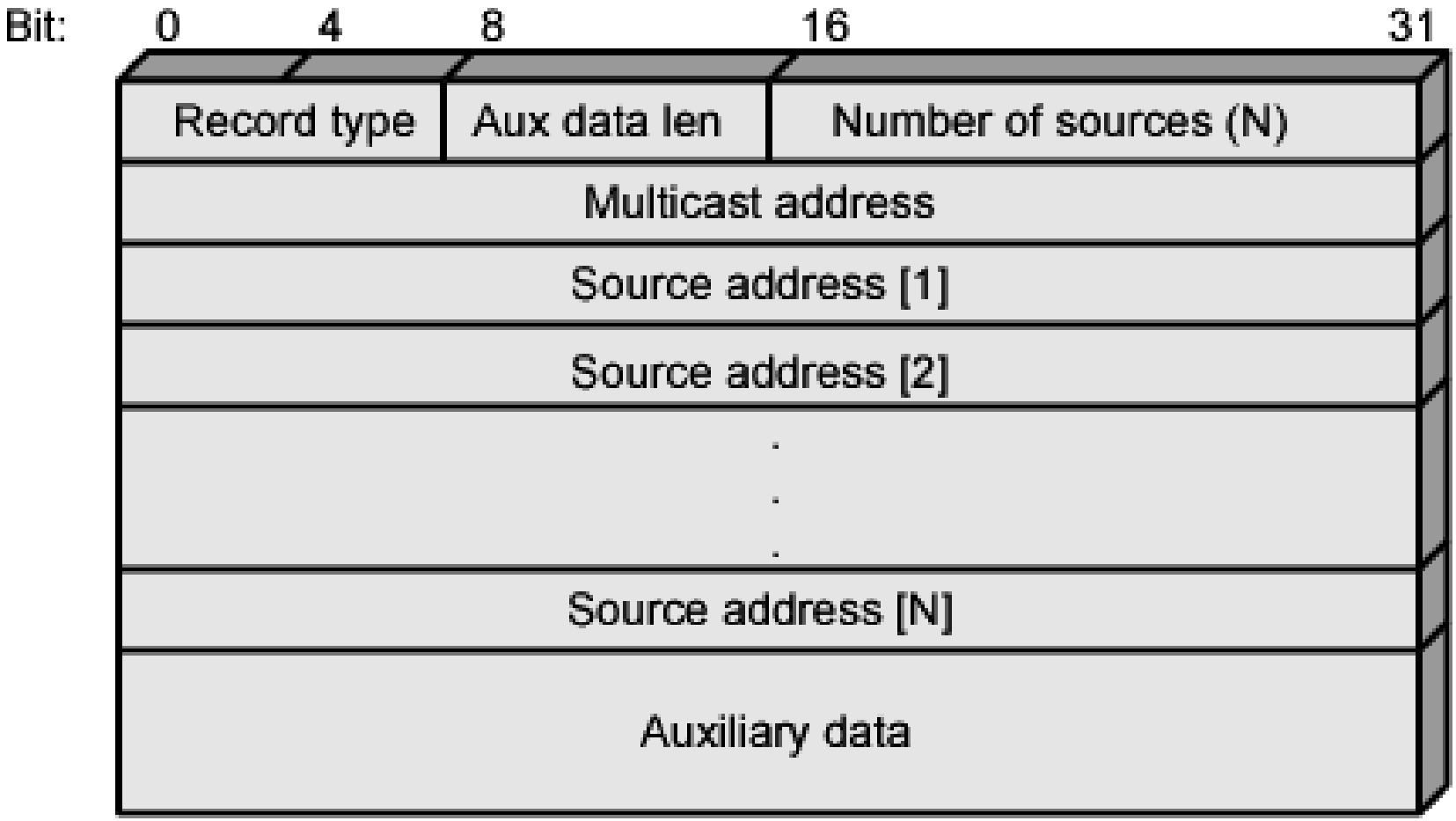


(b) Membership report message

Membership Reports

- Type
- Checksum
- Number of Group Records
- Group Records
  - One 32-bit unicast address per source

IGMP Message Formats
Group Record



(c) Group record

Group Record

- Record Type
  - —See later
- Aux Data Length
  - —In 32-bit words
- Number of Sources
- Multicast Address
- Source Addresses
  - —One 32-bit unicast address per source
- Auxiliary Data
  - —Currently, no auxiliary data values defined

IGMP Operation - Joining

- Host using IGMP wants to make itself known as group member to other hosts and routers on LAN
- IGMPv3 can signal group membership with filtering capabilities with respect to sources
  - EXCLUDE mode – all group members except those listed
  - INCLUDE mode – Only from group members listed
- To join group, host sends IGMP membership report message
  - Address field multicast address of group
  - Sent in IP datagram with Group Address field of IGMP message and Destination Address encapsulating IP header same
  - Current members of group will receive learn of new member
  - Routers listen to all IP multicast addresses to hear all reports

IGMP Operation – Keeping Lists Valid

- Routers periodically issue IGMP general query message
  - In datagram with all-hosts multicast address
  - Hosts that wish to remain in groups must read datagrams with this all-hosts address
  - Hosts respond with report message for each group to which it claims membership
- Router does not need to know every host in a group
  - Needs to know at least one group member still active
  - Each host in group sets timer with random delay
  - Host that hears another claim membership cancels own report
  - If timer expires, host sends report
  - Only one member of each group reports to router

IGMP Operation - Leaving

- Host leaves group, by sending leave group message to all-routers static multicast address
- Send membership report message with EXCLUDE option and null list of source addresses
- Router determine if there are any remaining group members using group-specific query message

Group Membership with IPv6

- IGMP defined for IPv4
  - —Uses 32-bit addresses
- IPv6 internets need functionality
- IGMP functions incorporated into Internet Control Message Protocol version 6 (ICMPv6)
  - —ICMPv6 includes all of functionality of ICMPv4 and IGMP
- ICMPv6 includes group-membership query and group-membership report message
  - —Used in the same fashion as in IGMP

Routing Protocols

- Routing Information
  - —About topology and delays in the internet
- Routing Algorithm
  - —Used to make routing decisions based on information

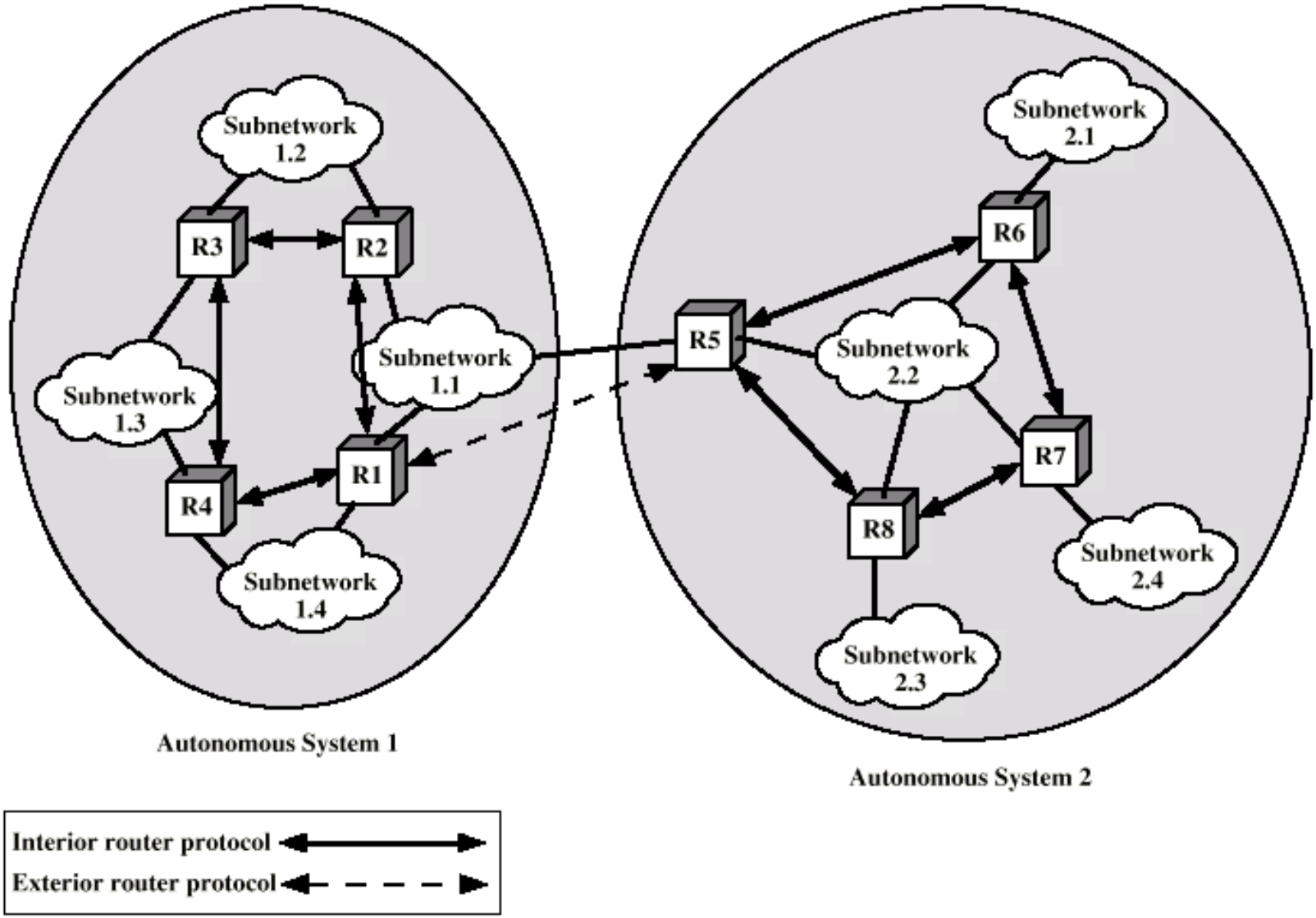
Autonomous Systems (AS)

- Group of routers
- Exchange information
- Common routing protocol
- Set of routers and networks managed by signle organization
- A connected network
  - —There is at least one route between any pair of nodes

Interior Router Protocol (IRP) Exterior Routing Protocol (ERP)

- Passes routing information between routers within AS
- May be more than one AS in internet
- Routing algorithms and tables may differ between different AS
- Routers need some info about networks outside their AS
- Used exterior router protocol (ERP)
- IRP needs detailed model
- ERP supports summary information on reachability

Application of IRP and ERP



Approaches to Routing – Distance-vector

- Each node (router or host) exchange information with neighboring nodes
  - Neighbors are both directly connected to same network
- First generation routing algorithm for ARPANET
- Node maintains vector of link costs for each directly attached network and distance and next-hop vectors for each destination
- Used by Routing Information Protocol (RIP)
- Requires transmission of lots of information by each router
  - Distance vector to all neighbors
  - Contains estimated path cost to all networks in configuration
  - Changes take long time to propagate

Approaches to Routing – Link-state

- Designed to overcome drawbacks of distance-vector
- When router initialized, it determines link cost on each interface
- Advertises set of link costs to all other routers in topology
  - Not just neighboring routers
- From then on, monitor link costs
  - If significant change, router advertises new set of link costs
- Each router can construct topology of entire configuration
  - Can calculate shortest path to each destination network
- Router constructs routing table, listing first hop to each destination
- Router does not use distributed routing algorithm
  - Use any routing algorithm to determine shortest paths
  - In practice, Dijkstra's algorithm
- Open shortest path first (OSPF) protocol uses link-state routing.
- Also second generation routing algorithm for ARPANET

Exterior Router Protocols – Not Distance-vector

- Link-state and distance-vector not effective for exterior router protocol
- Distance-vector assumes routers share common distance metric
- ASs may have different priorities
  - —May have restrictions that prohibit use of certain other AS
  - —Distance-vector gives no information about ASs visited on route

Exterior Router Protocols – Not Link-state

- Different ASs may use different metrics and have different restrictions
  - —Impossible to perform a consistent routing algorithm.
- Flooding of link state information to all routers unmanageable

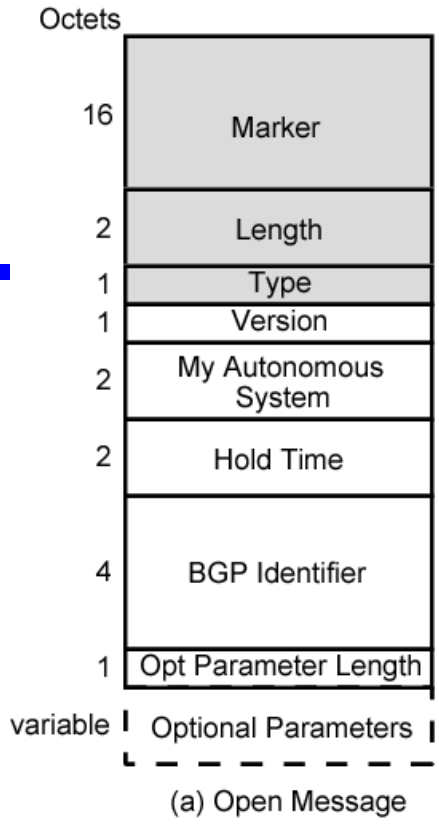
Exterior Router Protocols – Path-vector

- Dispense with routing metrics
- Provide information about which networks can be reached by a given router and ASs crossed to get there
  - Does not include distance or cost estimate
- Each block of information lists all ASs visited on this route
  - Enables router to perform policy routing
  - E.g. avoid path to avoid transiting particular AS
  - E.g. link speed, capacity, tendency to become congested, and overall quality of operation, security
  - E.g. minimizing number of transit ASs

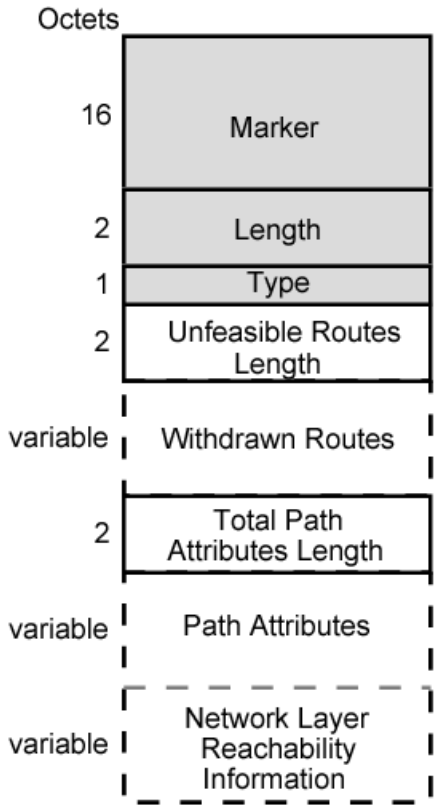
Border Gateway Protocol (BGP)

- For use with TCP/IP internets
- Preferred EGP of the Internet
- Messages sent over TCP connections
  - Open
  - Update
  - Keep alive
  - Notification
- Procedures
  - Neighbor acquisition
  - Neighbor reachability
  - Network reachability

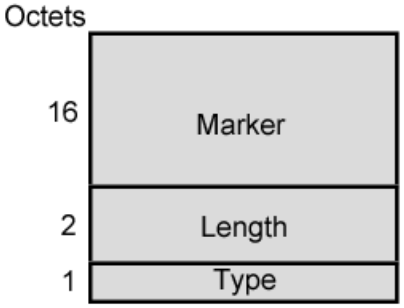
BGP Messages



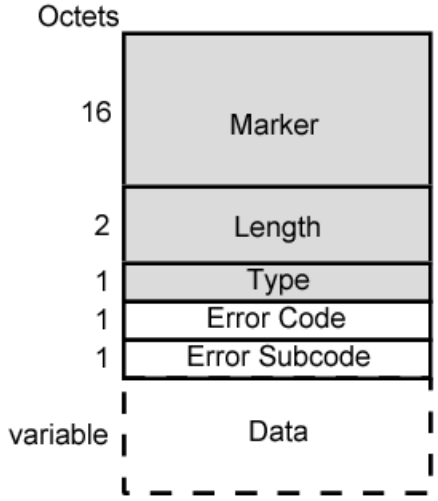
(a) Open Message



(b) Update Message



(c) Keepalive Message



(d) Notification Message

BGP Procedure

- Open TCP connection
- Send Open message
  - —Includes proposed hold time
- Receiver selects minimum of its hold time and that sent
  - —Max time between Keep alive and/or update messages

Message Types

- Keep Alive
  - To tell other routers that this router is still here
- Update
  - Info about single routes through internet
  - List of routes being withdrawn
  - Includes path info
    - Origin (IGP or EGP)
    - AS_Path (list of AS traversed)
    - Next_hop (IP address of boarder router)
    - Multi_Exit_Disc (Info about routers internal to AS)
    - Local_pref (Inform other routers within AS)
    - Atomic_Aggregate, Aggregator (Uses address tree structure to reduce amount of info needed)

Uses of AS_Path and Next_Hop

- AS_Path
  - —Enables routing policy
    - Avoid a particular AS
    - Security
    - Performance
    - Quality
    - Number of AS crossed
- Next_Hop
  - —Only a few routers implement BGP
    - Responsible for informing outside routers of routes to other networks in AS

Notification Message

- Message header error
  - Authentication and syntax
- Open message error
  - Syntax and option not recognized
  - Unacceptable hold time
- Update message error
  - Syntax and validity errors
- Hold time expired
  - Connection is closed
- Finite state machine error
- Cease
  - Used to close a connection when there is no error

BGP Routing Information Exchange

- Within AS, router builds topology picture using IGP
- Router issues Update message to other routers outside AS using BGP
- These routers exchange info with other routers in other AS
- Routers must then decide best routes

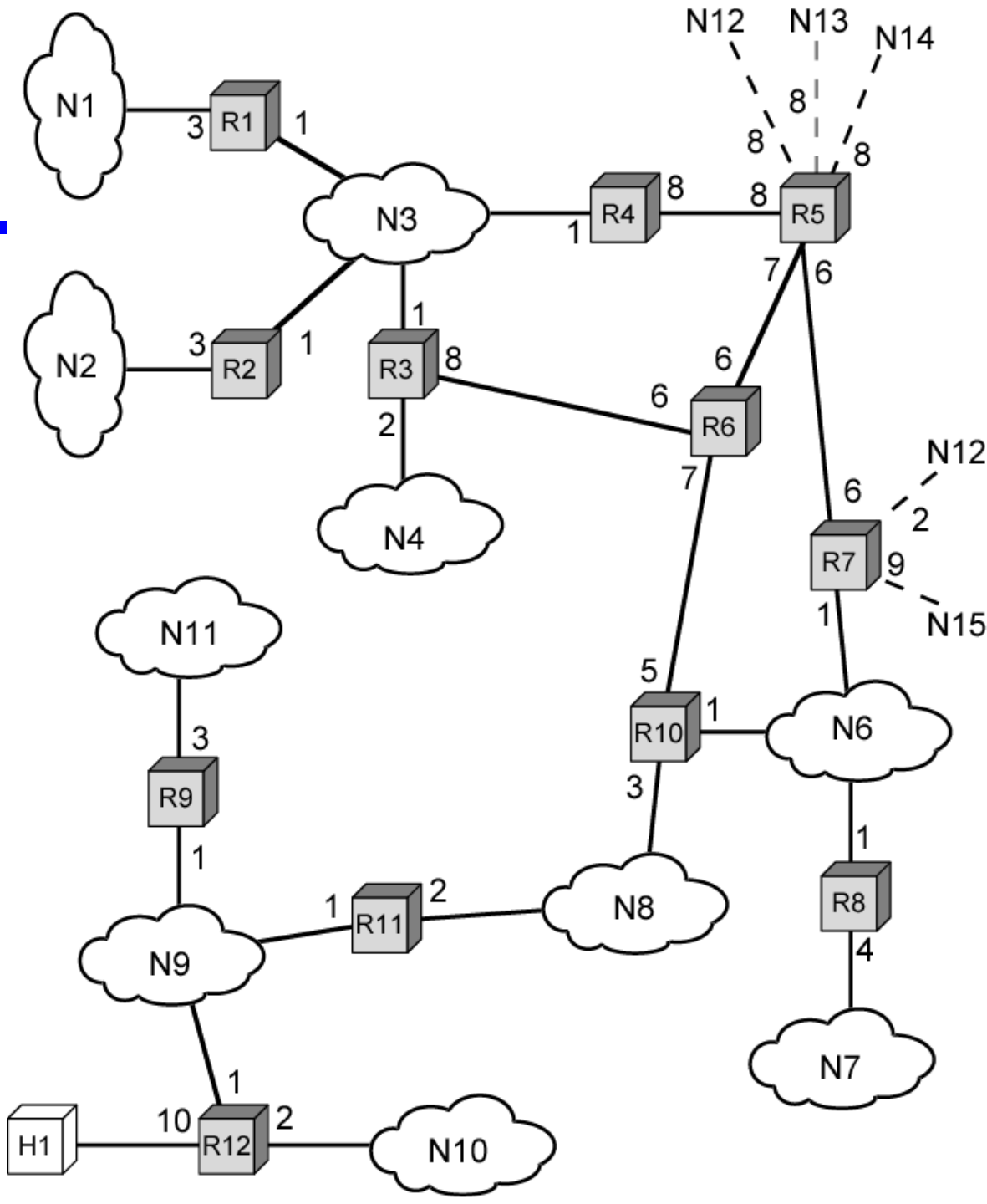
Open Shortest Path First (1)

- OSPF
- IGP of Internet
- Replaced Routing Information Protocol (RIP)
- Uses Link State Routing Algorithm
  - —Each router keeps list of state of local links to network
  - —Transmits update state info
  - —Little traffic as messages are small and not sent often
  - —RFC 2328
- Route computed on least cost based on user cost metric

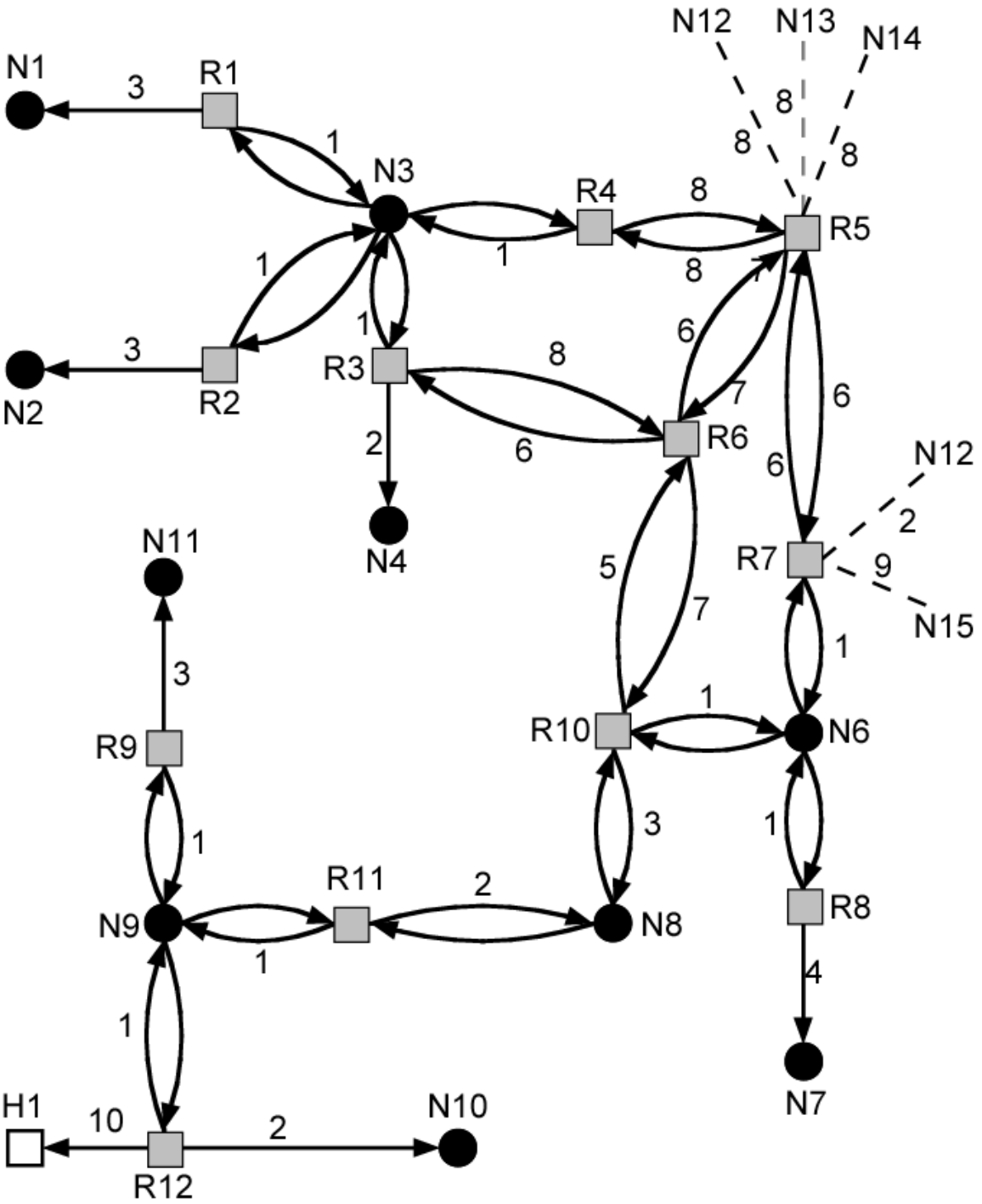
Open Shortest Path First (2)

- Topology stored as directed graph
- Vertices or nodes
  - —Router
  - —Network
    - Transit
    - Stub
- Edges
  - —Graph edge
    - Connect two router
    - Connect router to network

Sample AS


N1
R1
N2
R2
N3
R3
N4
R4
R5
N12
N13
N14
R6
N12
R7
N15
N11
R9
R10
N6
N8
R8
N9
R11
N7
H1
R12
N10

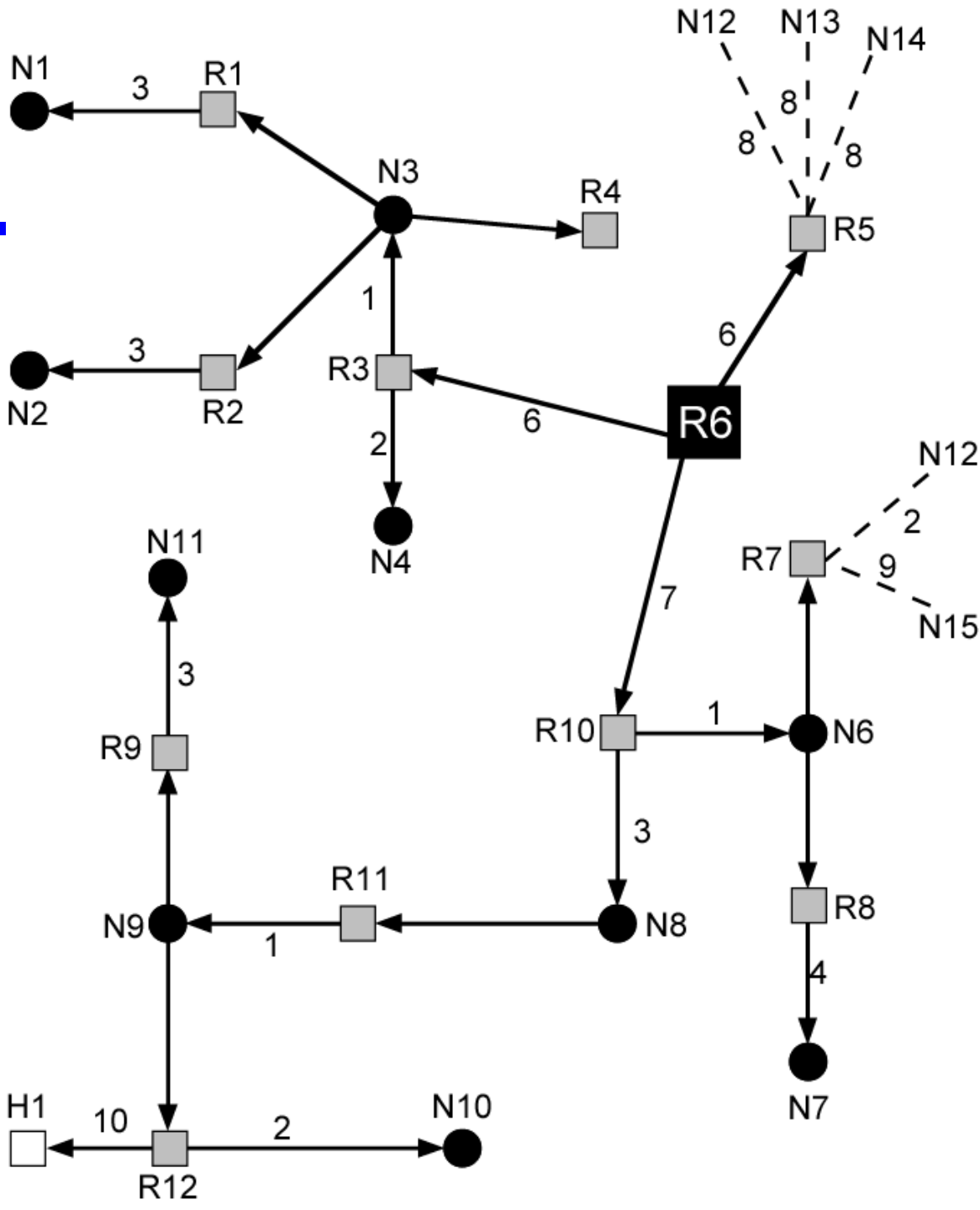
Directed Graph of AS



Operation

- Dijkstra's algorithm used to find least cost path to all other networks
- Next hop used in routing packets

SPF Tree for Router 6



Integrates Services Architecture

- Changes in traffic demands require variety of quality of service
- Internet phone, multimedia, multicast
- New functionality required in routers
- New means of requesting QoS
- ISA
- RFC 1633

Internet Traffic

- Elastic
  - —Can cope with wide changes in delay and/or throughput
    - FTP sensitive to throughput
    - E-Mail insensitive to delay
    - Network Management sensitive to delay in times of heavy congestion
    - Web sensitive to delay
- Inelastic
  - —Does not easily adapt to variations
  - —e.g. real time traffic

Requirements for Inelastic Traffic

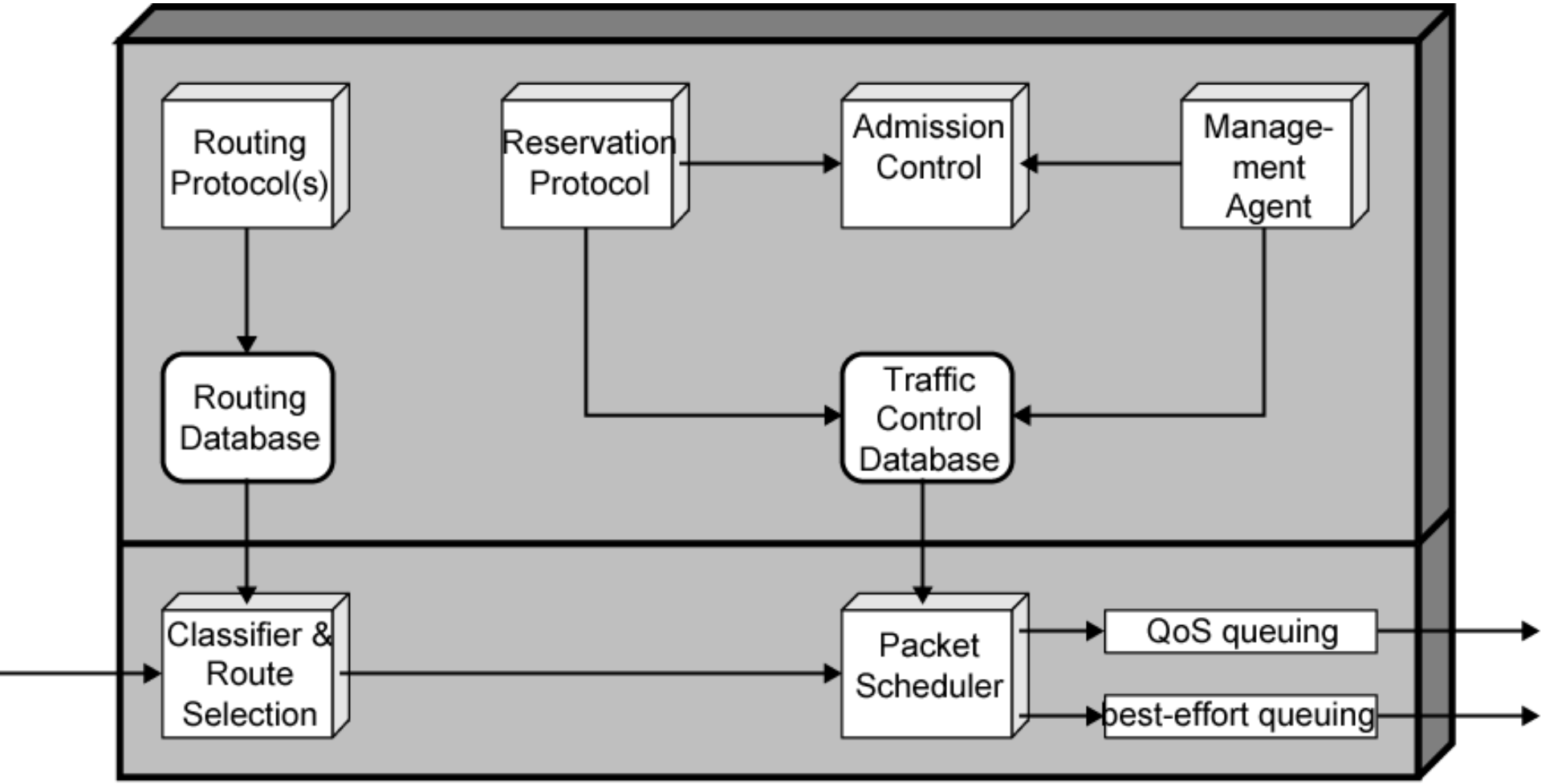
- Throughput
- Delay
- Jitter
  - Delay variation
- Packet loss

- Require preferential treatment for certain types of traffic
- Require elastic traffic to be supported as well

ISA Approach

- Congestion controlled by
  - —Routing algorithms
  - —Packet discard
- Associate each packet with a flow
  - —Unidirectional
  - —Can be multicast
- Admission Control
- Routing Algorithm
- Queuing discipline
- Discard policy

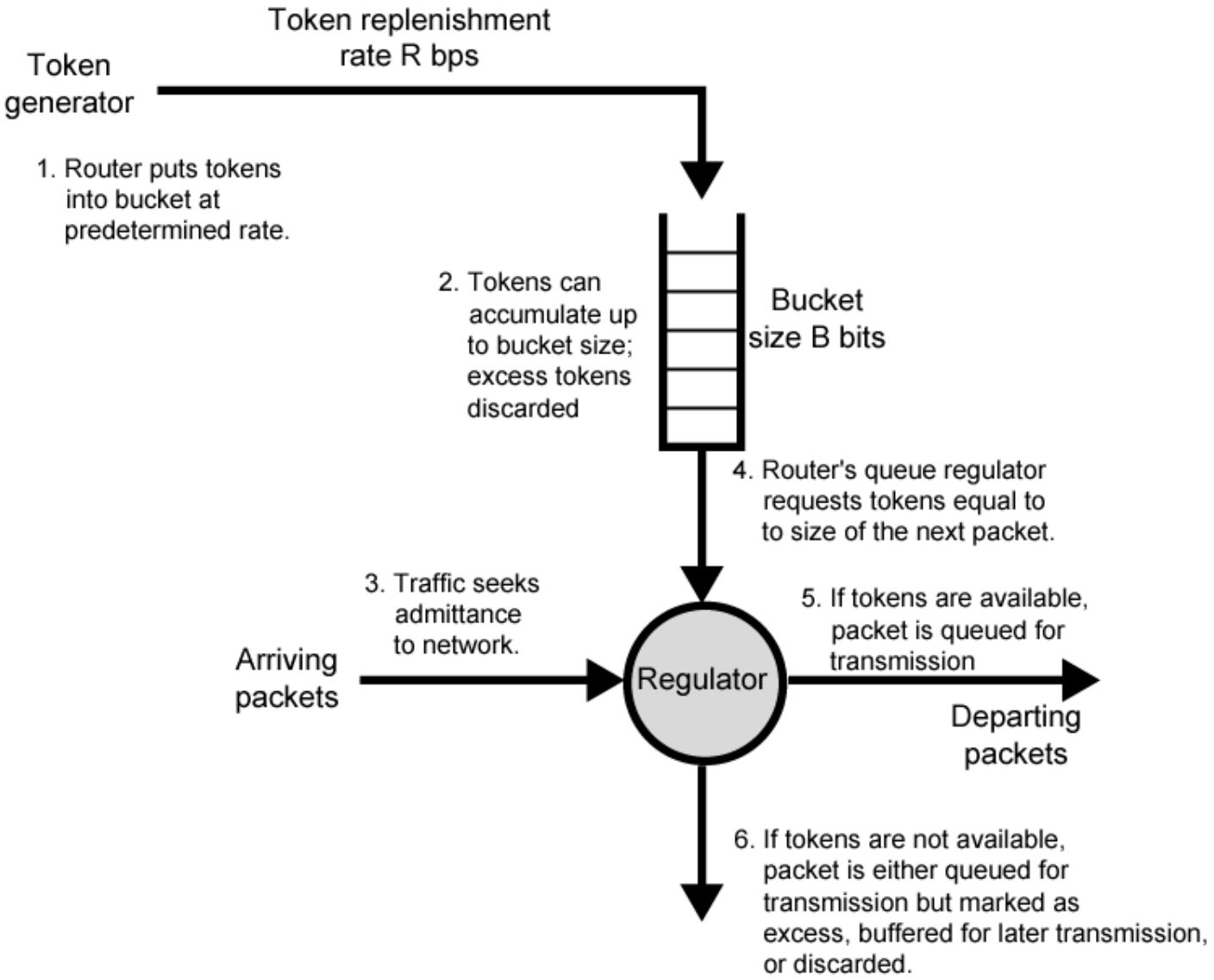
ISA in Router



Token Bucket Traffic Specification

- Token replenishment rate *R*
  - —Continually sustainable data rate
- Bucket size *B*
  - —Amount that data rate can exceed R for short period
  - —During time period *T* amount of data sent can not exceed *RT + B*

Token Bucket Scheme



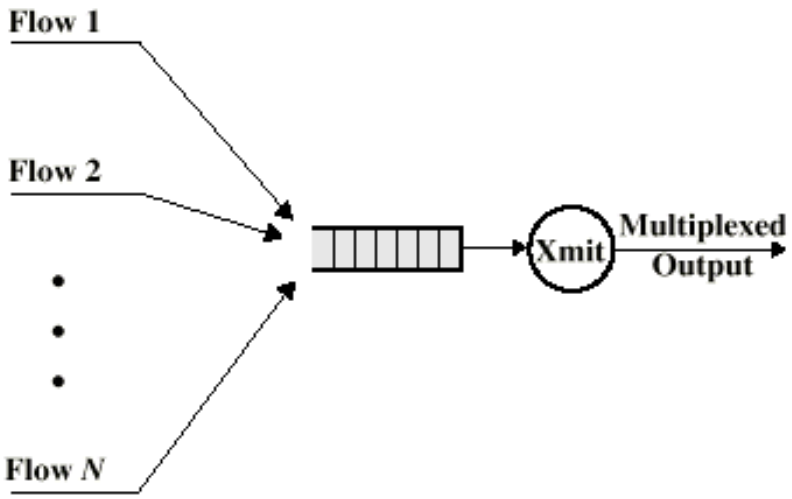
ISA Services

- Guaranteed
  - —Assured data rate
  - —Upper bound on queuing delay
  - —No queuing loss
  - —Real time playback
- Controlled load
  - —Approximates behavior to best efforts on unloaded network
  - —No specific upper bound on queuing delay
  - —Very high delivery success
- Best Effort

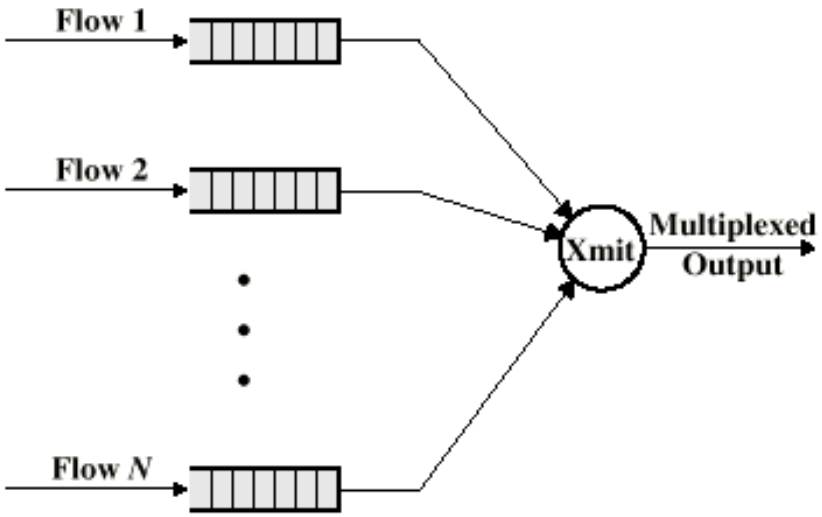
Queuing Discipline

- Traditionally FIFO
  - —No special treatment for high priority flow packets
  - —Large packet can hold up smaller packets
  - —Greedy connection can crowd out less greedy connection
- Fair queuing
  - —Queue maintained at each output port
  - —Packet placed in queue for its flow
  - —Round robin servicing
  - —Skip empty queues
  - —Can have weighted fair queuing

FIFO and Fair Queue



(a) FIFO Queuing



(b) Fair Queuing

Resource Reservation: RSVP

- RFC 2205
- Unicast applications can reserve resources in routers to meet QoS
- If router can not meet request, application informed
- Multicast is more demanding
- May be reduced
  - Some members of group may not require delivery from particular source over given time
    - e.g. selection of one from a number of "channels"
  - Some group members may only be able to handle a portion of the transmission

Soft State

- Set of state info in router that expires unless refreshed
- Applications must periodically renew requests during transmission

RSVP Characteristics

- Unicast and Multicast
- Simplex
- Receiver initiated reservation
- Maintain soft state in the internet
- Provide different reservation styles
- Transparent operation through non-RSVP routers
- Support for IPv4 and IPv6

Differentiated Services

- Provide simple, easy to implement, low overhead tool to support range of network services differentiated on basis of performance
- IP Packets labeled for differing QoS using existing IPv4 Type of Service or IPv6 Traffic class
- Service level agreement established between provider and customer prior to use of DS
- Built in aggregation
  - Good scaling to larger networks and loads
- Implemented by queuing and forwarding based on DS octet
  - No state info on packet flows stored

DS Services

- Defined within DS domain
  - —Contiguous portion of internet over which consistent set of DS policies are administered
  - —Typically under control of one organization
  - —Defined by service level agreements (SLA)

SLA Parameters

- Detailed service performance
  - —Expected throughput
  - —Drop probability
  - —Latency
- Constraints on ingress and egress points
- Traffic profiles
  - —e.g. token bucket parameters
- Disposition of traffic in excess of profile

Example Services

- Level A - low latency
- Level B - low loss
- Level C - 90% of traffic < 50ms latency
- Level D - 95% in profile traffic delivered
- Level E - allotted twice bandwidth of level F traffic
- Traffic with drop precedence X higher probability of delivery than that of Y

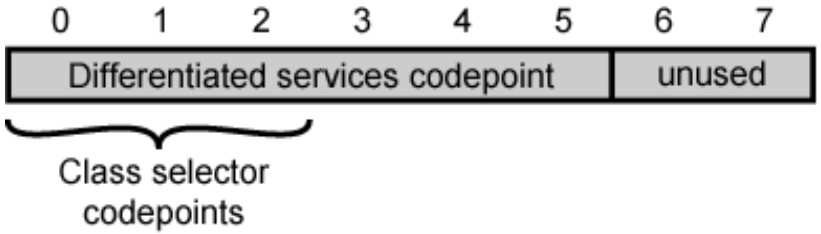
DS Octet - Code Pools

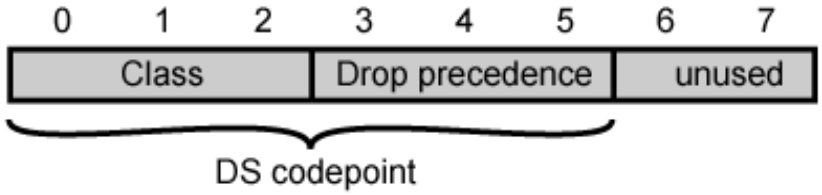
- Leftmost 6 bits used
- 3 pools of code points
- xxxxx0
  - —assignment as standards
- xxxx11
  - —experimental or local use
- xxxx01
  - —experimental or local but may be allocated for standards in future

DS Octet - Precedence Field

- Routing selection
- Network service
- Queuing discipline

DS Field





000000	Default behavior
001000	Class selector behaviors
010000	
011000	
100000	
101000	
110000	
111000	
101110	Expedited forwarding (EF) behavior

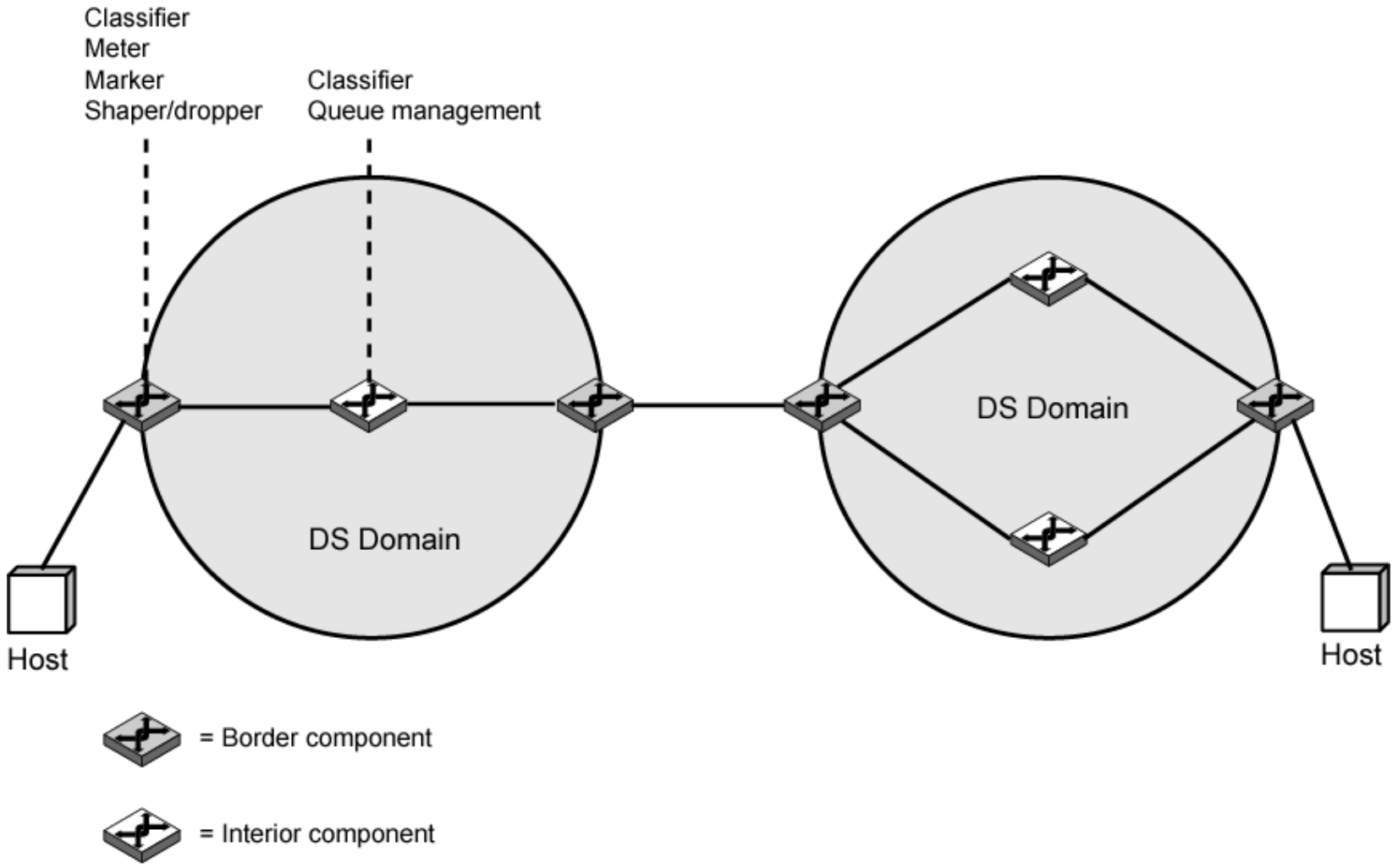
Increasing priority

(a) DS Field

Class		Drop Precedence	
100	Class 4 - best service	010	Low - most important
011	Class 3	100	Medium
010	Class2	110	High - least important
001	Class 1		

(b) Codepoints for assured forwarding PHB

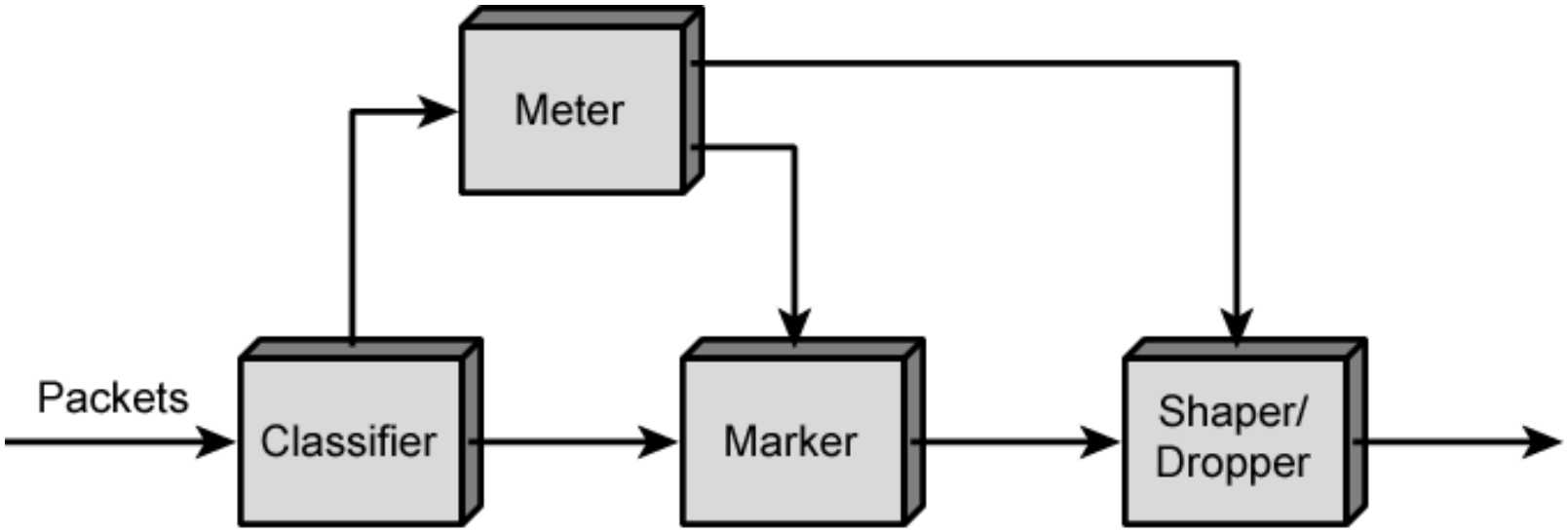
DS Domains


Classifier
Meter
Marker
Shaper/dropper
Classifier
Queue management
DS Domain
DS Domain
Host
Host
= Border component
= Interior component

DS Configuration and Operation

- Within domain, interpretation of DS code points is uniform
- Routers in domain are boundary nodes or interior nodes
- Traffic conditioning functions
  - Classifier
  - Meter
  - Marker
  - Shaper
  - Dropper

DS Traffic Conditioner



Per Hop Behavior – Expedited Forwarding

- Specific PHBs defined,
  - Associated with specific differentiated services
- RFC 3246 defines expedited forwarding (EF) PHB
  - Support for premium service
  - Low-loss, low-delay, low-jitter, assured bandwidth, end-to-end service through DS domains
  - Appears to endpoints as point-to-point connection or leased line
- Difficult in internet or packet-switching network
  - Queues (buffers) at each node, or router
  - Results in loss, delays, and jitter
  - Unless internet grossly oversized, care needed in handling premium service traffic

Expedited Forwarding Requirements

- Configuring nodes so traffic aggregate has minimum departure rate
- Conditioning aggregate (via policing and shaping) so that arrival rate less than node's configured minimum departure rate
- EF PHB provides first
- Network boundary conditioners provide second
- Border nodes control traffic aggregate
  - Limit characteristics (rate, burstiness) to predefined level
- Interior nodes treat traffic so no queuing effects
- No specific queuing policy at interior nodes in RFC 3246
- Simple priority scheme could achieve it
    - EF traffic given absolute priority
    - EF traffic must not overwhelm interior node
    - Packet flows for other PHB traffic disrupted

Assured Forwarding PHB

- Provide service superior to best-effort
- Not require reservation of resources
- Not require detailed discrimination among flows from different users
- Based on explicit allocation
  - Users offered choice of classes of service
    - Each class describes different traffic profile
    - Aggregate data rate and burstiness
  - Traffic monitored at boundary node
    - Each packet marked in or out of profile
  - Inside network, no separation of traffic from different users or classes
    - Only distinction being whether packet marked in or out
  - When congested, out packets are dropped before in packets
  - Different users will see different levels of service
    - Have different quantities of in packets in service queues

Advantages of Assured Forwarding PHB

- Simplicity
  - —Very little work required by internal nodes
  - —Marking of traffic at boundary nodes based on traffic profiles provides different levels of service to different classes
- C.f. ATM

AF PHB RFC 2597 (1)

- Four AF classes defined
  - Four distinct traffic profiles
- Within each class, packets marked by customer or service provider
  - Three drop precedence values
    - Determines relative importance of packet within AF class
- Simpler than resource reservation
- Flexible
- Within interior DS node, traffic from different classes treated separately
  - Different amounts of resources (buffer space, data rate)

AF PHB RFC 2597 (2)

- Within class, packets handled based on drop precedence
- Level of forwarding assurance depends on:
  - How much forwarding resources allocated to AF class that the packet belongs to
  - Current load of class
  - If congested within the class, drop precedence of packet
- RFC 2597 does not mandate mechanisms at the interior nodes to manage AF traffic
  - References RED algorithm

Required Reading

- Stallings chapter 19
- Comer, S. Internetworking with TCP/IP, volume 1, Prentice-Hall
- All RFCs mentioned plus any others connected with these topics
- Loads of Web sites on TCP/IP, routing protocols etc.